Chapter 20:

The CrimeStat Regression Module[1]

Ned Levine
Ned Levine & Associates
Houston, TX

Dominique Lord
Zachry Dept. of Civil Engineering
Texas A & M University
College Station, TX

Byung-Jung Park
Korea Transport Institute
Goyang, South Korea

Srinivas Geedipally
Texas Transportation Institute
Arlington, TX

Haiyan Teng
Houston, TX

Li Sheng
Houston, TX

Ian Cahill
Cahill Software
Edmonton, AB

[1] The regression chapters were the result of the effort of many persons. The maximum likelihood routines were produced by Ian Cahill of Cahill Software in Edmonton, Alberta as part of his MLE++ software package. We are grateful to him for providing these routines and for conducting quality control tests. Dr. Shaw-pin Miaou of College Station, TX designed the MCMC algorithm for the Poisson-Gamma-CAR model. Dr. Byung-Jung Park modified the algorithm to incorporate Poisson-Gamma-SAR, Poisson-Lognormal-CAR/SAR, and the MCMC binomial-CAR/SAR models. Dr. Srinivas Geeidpally added the MCMC Normal-CAR/SAR models. Dr. Dominique Lord of Texas A & M University provided technical consulting on the dispersion parameters in these models. Dr. Ned Levine developed the block sampling scheme and provided overall project management. Ms. Haiyan Teng and Dr. Li Sheng programmed the routines and added numerous technical improvements to the algorithms. We are also grateful to Dr. Richard Block of Loyola University in Chicago (IL) for testing the MCMC and MLE routines.

Table of Contents

Table of Contents (continued)

Chapter 20:
The CrimeStat Regression Module

We now describe the *CrimeStat* regression module. There are two pages in the module. Regression I allows the testing of a model while Regression II allows a prediction to be made based on an already-estimated model. Figure 20.1 displays the Regression I page.

Regression I Module

Types of Regression Models

In the current version, 18 possible regression models are available with several options for each of these:

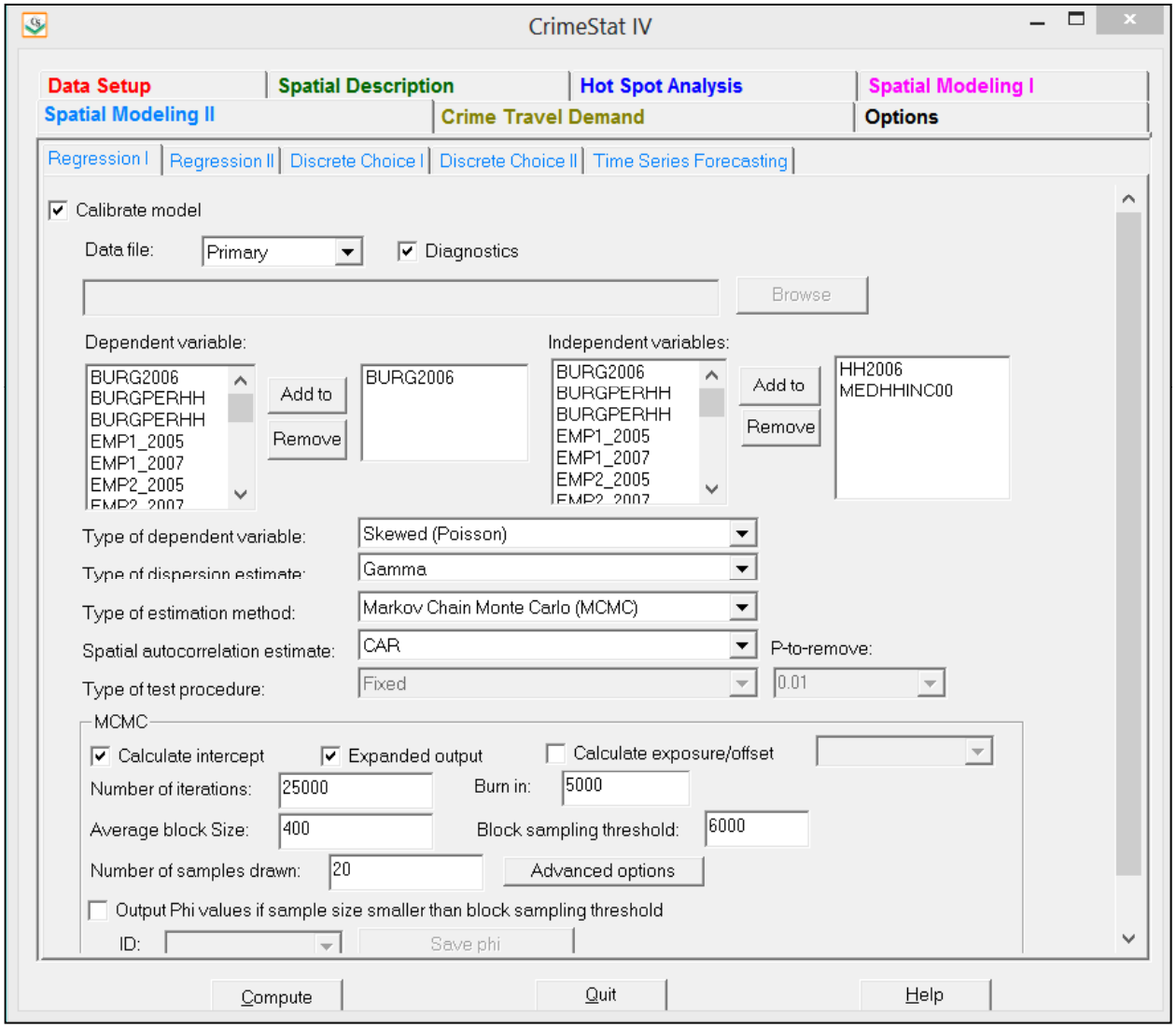
MLE Normal (OLS)
MCMC Normal
MCMC Normal-CAR
MCMC Normal-SAR
MLE Poisson
MLE Poisson with linear dispersion correction (NB1)
MLE Poisson-Gamma (NB2)
MCMC Poisson-Gamma (NB2)
MCMC Poisson-Gamma-CAR
MCMC Poisson-Gamma-SAR
MCMC Poisson-Lognormal
MCMC Poisson-Lognormal-CAR
MCMC Poisson-Lognormal-SAR
MLE Binomial Logit
MLE Binomial Probit
MCMC Binomial Logit
MCMC Binomial Logit-CAR
MCMC Binomial Logit-SAR

In addition, each of the 12 MCMC models can be run with an exposure (offset) variable used to define the population 'at risk' allowing a total of 30 possible regression models to be run.

There are two pages in the module. The Regression I page allows the testing of a model while the Regression II page allows a prediction to be made based on an already-estimated

Figure 20.1:

Regression Modeling I Setup Screen



model. Also, since the Regression I module and Trip Generation module in the Crime Travel Demand Model duplicate regression functions, only one of these can be run at a time.

Input Data Set

The data set for the regression module is the Primary File data set. The coordinate system and distance units are also the same. The routine will not work unless the Primary File has X/Y coordinates.

Dependent Variable

To start loading the module, click on the 'Calibrate model' tab. A list of variables from the Primary File is displayed. There is a box for defining the dependent variable. The user must choose one dependent variable. A keystroke trick is to click on the first letter of the variable that will be the dependent variable and the routine will go to the first variable with that letter.

Independent Variables

There is another box for defining the independent variables. The user must choose one or more independent variables. In the routine, there is no limit to the number. Keep in mind that the variables are output in the same order as specified in the dialogue so a user might want to think how these should be displayed.

Type of Dependent Variable

There are five options that must be defined. The first is the type of dependent variable: Skewed (Poisson), Normal (OLS), Binomial probit, or Binomial logit (logistic). The default is a Poisson.

Type of Dispersion Estimate

The second model decision is the type of dispersion estimate to be used. The choices are Gamma, Poisson, Lognormal, and Poisson with linear correction. For the MLE models, only Gamma, Poisson and Poisson with linear correction are available while for the MCMC models, only Gamma and Lognormal are available. The default is Gamma. For the MLE Normal (OLS) and MCMC Normal-CAR/SAR models, the dispersion is automatically normal. For the binomial logit or binomial probit, the dispersion is automatically binomial.

Type of Estimation Method

The third option is the type of estimation method to be used: Maximum Likelihood (MLE) or Markov Chain Monte Carlo (MCMC). The default is MLE. These methods were discussed in Chapters 15 and 17 and in appendices B and C.

Spatial Autocorrelation Estimate

Fourth, if the user accepts an MCMC algorithm, then a fourth decision is whether to run a spatial autocorrelation estimate along with it (a Conditional Autoregressive function - CAR, or a Simultaneous Autoregressive function - SAR). The MCMC Poisson-Gamma, MCMC Poisson-Lognormal, and MCMC Logit functions can be run with a spatial autocorrelation parameter.

> Note that the CAR model runs quite quickly whereas the SAR model runs very slowly. Unless the data set is small or a SAR model is absolutely essential, we recommend using a CAR function for the spatial regression models.

Type of Test Procedure

The fifth, and last model decision, is whether to run a fixed model or a backward elimination *stepwise* procedure (only with the normal or MLE models). A fixed model includes all selected independent variables in the regression whereas a backward elimination model starts with all selected variables in the model but proceeds to drop variables that fail the P-to-remove test, one at a time. Any variable that has a significance level in excess of the P-to-remove value is dropped from the equation.

If the fixed model is chosen, then all independent variables will be regressed simultaneously. However, if the stepwise backward elimination procedure is selected, the user must define a *p-to-remove* value. The choices are: 0.1, 0.05, 0.01, and 0.001. The default is 0.01. Traditionally, 0.05 is used as a minimal threshold for significance. We put in 0.01 as the default to make the model stricter; with the large datasets that typically occur in police departments, the less strict 0.05 criterion would not exclude many independent variables. But, the user can certainly use 0.05 instead.

MCMC Choices

If the user chooses the MCMC algorithm, then nine *additional* decisions have to be made.

Number of Iterations

The first MCMC decision is the number of iterations to be run. The default is 25,000. The number should be sufficient to produce reliable estimates of the parameters. Check the MC Error/Standard deviation ratio and the G-R statistic to be sure these are below 1.05 and 1.20 respectively.

'Burn in' Iterations

The second MCMC decision is the number of initial iterations that will be dropped from the final distribution (the 'burn in' period). The default is 5,000. The number of 'burn in' iterations should be sufficient for the algorithm to reach an equilibrium state and produce reliable estimates of the parameters. Check the MC Error/Standard deviation ratio and the G-R statistic to be sure these are below 1.05 and 1.20 respectively.

Block Sampling Threshold

The third MCMC decision is whether to run all the records through the MCMC algorithm or whether to draw block samples. This is called the *Block Sampling Threshold*. The algorithm will be run on all cases unless the number of records exceeds the number specified in the block sampling threshold. The default threshold is 6,000 cases. If the number of cases exceeds the threshold, then the block sampling method is used (see below).

Note that if you raise the run the block sampling threshold for more cases, calculating time will increase substantially. For the non-spatial Poisson-Gamma model, the increase is linear. However, for the spatial Poisson-Gamma model, the increase is exponential. Further, we have found that we cannot calculate the spatial model for more than about 6,000 cases. In short, the block sampling method must be used for spatial models with a large number of cases.

Average Block Size

The fourth MCMC decision is the number of cases to be drawn in each block sample if the total number of records is greater than the block sampling threshold. The default is 400 cases. Note that this is an average. Actual samples will vary in size. The output will display the expected sample size and the average sample size that was drawn.

Number of Samples Drawn

The fifth MCMC decision is the number of samples to be drawn if the total number of records is greater than the block sampling threshold. The default is 25. We have found that

reliable estimates can be obtained from 20 to 30 samples especially if the sequence converges quickly and even 10 samples can produce meaningful results. Obviously, the more samples that are drawn, the more reliable will be the final results. But, having more samples will not necessarily increase the precision beyond 30.

Calculate Intercept

The sixth MCMC decision is whether to run the model with or without an intercept (constant). The default is with an intercept estimated. To run the model without the intercept, uncheck the 'Calculate intercept' box.

Spatial Autocorrelation Estimate

The seventh MCMC decision is whether to run a spatial autocorrelation model. There are two alternative spatial autocorrelation functions that can be used, a *Conditional Autoregressive* (or CAR) or a *Simultaneous Autoregressive* (or SAR). These were defined in Chapter 19. The default is no spatial autocorrelation. Note that estimating the SAR function takes a long time, much longer than for the CAR model. Unless there is a reason for using the SAR, we recommend using the CAR for any spatial autocorrelation component.

Calculate Exposure/Offset

The eighth MCMC decision is whether to run a risk model. If the model is a risk or rate model, then an exposure (offset) variable needs to be defined. Check the 'Calculate exposure/offset' box and identify the variable that will be used as the exposure variable. The coefficient for this variable will automatically be 1.0.

Advanced Options

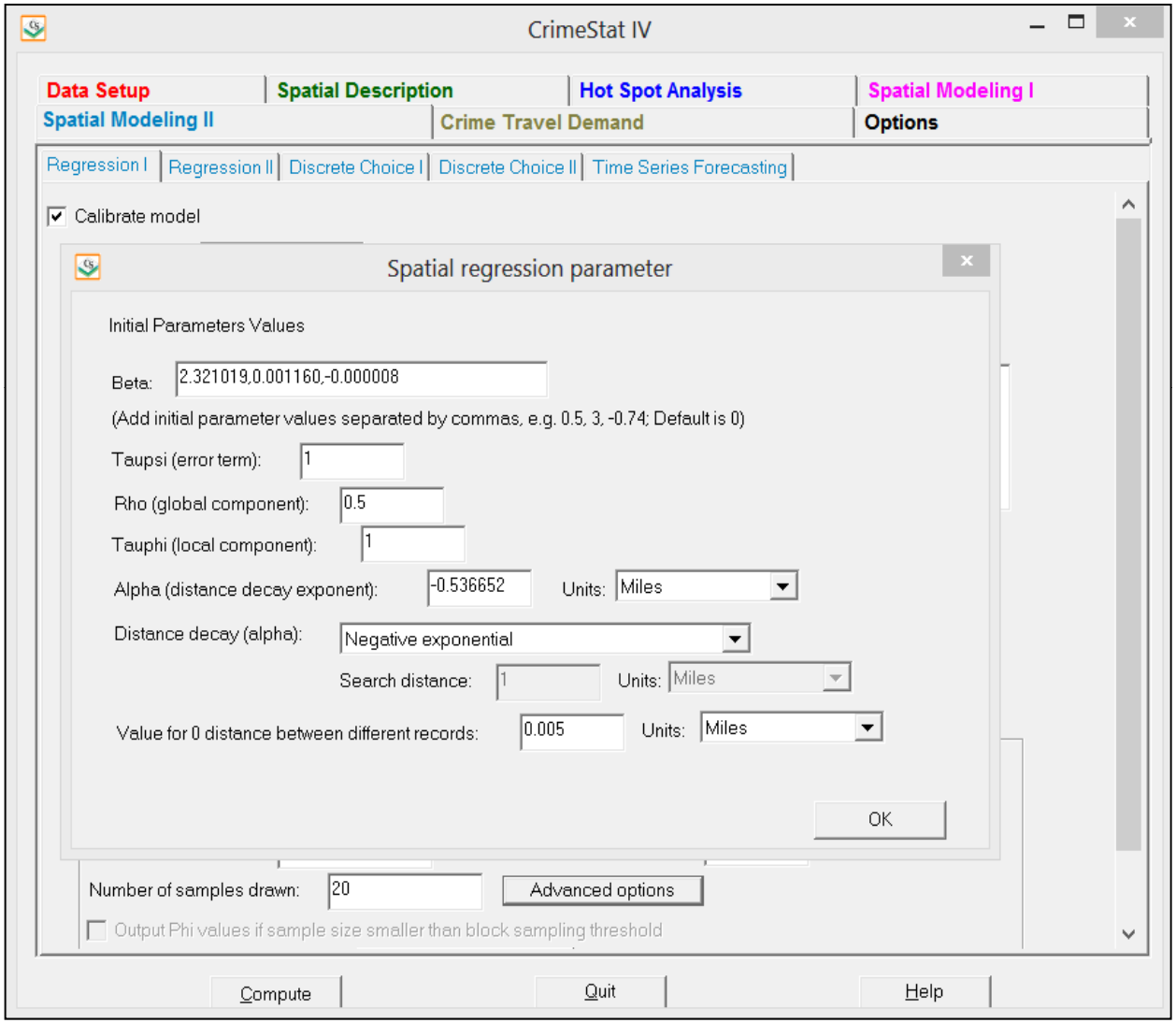
There is also a set of advanced options for the MCMC algorithm. Figure 20.2 displays the advanced options dialogue. We would suggest keeping the default values initially until you become very familiar with the routine.

Initial parameters values for Phi (φ)

The ninth, and last, MCMC decision is the prior values used for the different parameters being estimated. The MCMC algorithm requires an initial estimate for each parameter. There are default values that are used. For the beta coefficients (including the intercept), the default values are 0. This assumes that the coefficient is 'not significant' and has a large variance. It is frequently called a 'non-informative' prior. These are displayed as a blank screen for the Beta

Figure 20.2:

Advanced Options for MCMC Poisson-Gamma-CAR Model



box. However, estimates of the beta coefficients can be substituted for the assumed 0 coefficients. To do this, all independent variable coefficients plus the intercept (if used) must be listed in the order in which they appear in the model and must be separated by commas. Do not include the beta coefficients for the spatial autocorrelation, Φ_i, term (if used).

For example, suppose there are three independent variables. Thus, the model will have four coefficients (the intercept and the coefficients for each of three independent variables). Suppose a prior study had been done in which a Poisson-Gamma model was estimated as:

$$Y_i = e^{4.5+0.3X_{1i}-2.1X_{2i}+3.4X_{3i}} \quad (20.1)$$

The researcher wants to repeat this model but with a different data set and assumes that the model using the new data set will have coefficients similar to the earlier research. Thus, the following would be specified in the box for the betas under the advanced options:

4.5, 0.3, -2.1, 3.4 (20.2)

The routine will use these values for the initial estimates of the parameters before starting the MCMC process (with or without the block sampling method). The advantage is that the distribution will converge more quickly (assuming the model is appropriate for the new data set).

Rho (ρ) *and Tauphi* (τ_ϕ)

The spatial autocorrelation component, Φ, is made up of three separate sub-components, called Rho (ρ), Tauphi (τ_ϕ), and Alpha (α, see formula 19.5 in chapter 19). These are additive.

Rho is roughly a global component that applies to the entire data set. Tauphi is roughly a neighborhood component that applies to a sub-set of the data. Alpha is essentially a localized effect. The routine works by estimating values for Rho and Tauphi but uses a pre-defined value for Alpha. The default initial values for Rho and Tauphi are 0.5 and 1 respectively. The user can substitute alternative values for these parameters.

Alpha (α)

Alpha (α) is the exponent for the distance decay function in the spatial model. Essentially, the distance decay function defines the weight to be applied to the values of nearby records. The weight can be defined by one of three mathematical functions. First, the weight can be defined by a negative exponential function where:

$$Weight = e^{-\alpha d_{ij}} \quad (20.3)$$

where d_{ij} is the distance between observations and α is the value for alpha. It is automatically assumed that alpha will be negative whether the user puts in a minus sign or not. The user inputs the alpha value in this box.

Second, the weight can be defined by a restricted negative exponential whereby the negative exponential operates up to the specified search distance, whereupon the weight becomes 0 for greater distances:

Up to Search distance: $Weight = e^{-\alpha d_{ij}}$ for $d_{ij} \geq 0, d_{ij} \leq d_p$ (20.4)

Beyond search distance: 0 for $d_{ij} > d_p$ (20.5)

where d_p is the search distance. The coefficient for the linear component is assumed to be 1.0.

Third, the weight can be defined as a uniform value for all other observations within a specified search distance. This is a *contiguity* (or adjacency) measure. Essentially, all other observations have an equal weight within the search distance and 0 if they are greater than the search distance. The user inputs the search distance and units in this box.

For the negative exponential and restricted negative exponential functions, substitute the selected value for α in the alpha box.

Diagnostic test for reasonable alpha (α) value

The default function for the weight is a negative exponential with a default alpha value of -1 in miles. For many data sets, this will be a reasonable value. However, for other data sets, it will not.

Reasonable values for alpha with the negative exponential function are obtained with the following procedure:

1. Decide on the measurement units to be used to calculate alpha (miles, kilometers, feet, etc). The default is miles. *CrimeStat* will convert from the units defined for the Primary File input dataset to those specified by the user.

2. Calculate the nearest neighbor distance from the Nna routine on the Distance Analysis I page. These may have to be converted into units that were selected in step 1 above. For example, if the Nearest Neighbor distance is listed as 2000 feet, but the desired units for alpha are miles, convert 2000 feet to miles by dividing the 2000 by 5280.

3. Input the dependent variable as the Z (intensity) variable on the Primary File page.

4. Run the Moran Correlogram routine on this variable on the Spatial Autocorrelation page (under Spatial Description). By looking at the values and the graph, decide whether the distance decay in this variable is very 'sharp' (drops off quickly) or very 'shallow' (drops off slowly).

5. Define the appropriate weight for the nearest neighbor distance:

    a. Assume that the weight for an observation with itself (i.e., distance = 0) is 1.0.

    b. If the distance decay drops off sharply, then a low weight for nearby values should be given. Assume that any observations at the nearest neighbor distance will only have a weight of 0.5 with observations further away being even lower.

    c. If the distance decay drops off more slowly, then a higher weight for nearby values should be given. Assume that any observations at the nearest neighbor distance will have a weight of 0.9 with observations further away being lower but only slightly so.

    d. An intermediate value for the weight is to assume it to be 0.75.

6. A range of alpha values can be solved using these scenarios:

    a. For a sharp decay, alpha is given by:

$$\alpha = \frac{Ln(0.5)}{NN_{distance}} \tag{20.6}$$

    b. For a shallow distance decay, alpha is given by:

$$\alpha = \frac{Ln(0.9)}{NN_{distance}} \tag{20.7}$$

    c. For an intermediate decay, alpha is given by:

$$\alpha = \frac{Ln(0.75)}{NN_{distance}} \tag{20.8}$$

In all three equations, $NN_{distance}$ is the nearest neighbor distance.

These calculations will provide a range of appropriate values for α. The diagnostics routine automatically estimates these values as part of its output.

Value for 0 distance between records

The advanced options dialogue has a parameter for the minimum distance to be assumed between different records. If two records have the same X and Y coordinates (which could happen if the data are individual events, for example), then the distance between these records will be 0. This could cause unusual calculations in estimating spatial effects. Instead, it is more reliable to assume a slight difference in distance between all records. The default is 0.005 miles but the user can modify this (including substituting 0 for the minimal distance).

Output

The output depends on whether an MLE or an MCMC model has been run.

Maximum Likelihood (MLE) Model Output

The MLE routines (Normal, Poisson, Poisson with linear correction, MLE Poisson-Gamma, Binomial Probit, and MLE Binomial Logit/Logistic) produce a standard output which includes summary statistics and estimates for the individual coefficients.

MLE Summary Statistics

The summary statistics include:

Information About the Model

1. The data file
2. The dependent variable
3. The number of cases
4. The degrees of freedom (N – number of parameters estimated)
5. The type of regression model (Normal/OLS, Poisson, Poisson with linear correction, Poisson-Gamma, Binomial Logit)
6. The method of estimation (MLE)

Likelihood Statistics

7. Log-likelihood estimate, which is a negative number. For a set number of independent variables, the more negative the log-likelihood the better.

8. Log-likelihood per case. This divides the log-likelihood by the sample size (N). This indicates the average contribution to the log-likelihood of each observation. The more negative, the better.
9. Akaike Information Criterion (AIC) adjusts the log-likelihood for the degrees of freedom. The smaller the AIC, the better.
10. AIC per case. This divides the AIC statistic by the sample size (N). This indicates the average contribution to the AIC of each observation. The smaller, the better.
11. Bayesian Information Criterion (BIC), sometimes known as the Schwartz Criterion (SC), adjusts the log-likelihood for the degrees of freedom. The smaller the BIC, the better.
12. BIC per case. This divides the BIC/SC statistic by the sample size (N). This indicates the average contribution to the BIC/SC of each observation. The smaller, the better.
13. Deviance compares the log-likelihood of the model to the log-likelihood of a model that fits the data perfectly. A smaller deviance is better.
14. The probability value of the deviance based on a Chi-square test with *N-K-1* degrees of freedom where *K* is the number of independent variables.
15. Pearson Chi-square is a test of how closely the predicted model fits the data. A smaller Chi-square is better since it indicates the model fits the data well.
16. The probability value of the Pearson Chi-square based on a Chi-square test with *N-K-1* degrees of freedom where *K* is the number of independent variables.

Model Error Estimates

17. Mean Absolute Deviation (MAD). For a set number of independent variables, a smaller MAD is better.
18. Quartiles for the Mean Absolute Deviation. For any one quartile, smaller is better.
19. Mean Squared Predictive Error (MSPE). For a set number of independent variables, a smaller MSPE is better.
20. Quartiles for the Mean Squared Predictive Error. For any one quartile, smaller is better.
21. Squared multiple R (for linear model only). This is the percentage of the dependent variable accounted for by the independent variables.
22. Adjusted squared multiple R (for linear model only). This is the squared multiple R adjusted for degrees of freedom.

Dispersion Tests

23. Adjusted deviance. This is a measure of the difference between the observed and predicted values (the residual error) adjusted for degrees of freedom. The smaller the adjusted deviance, the better. A value greater than 1 indicates over-dispersion.
24. Probability of adjusted deviance. This is the Pearson Chi-square test with 1 degree of freedom.
25. Adjusted Pearson Chi-square. This is the Pearson Chi-square adjusted for degrees of freedom. The smaller the Pearson Chi-square, the better. A value greater than 1 indicates over-dispersion.
26. Probability of adjusted Pearson Chi-square. This is the Pearson Chi-square test with 1 degree of freedom.
27. Dispersion multiplier. This is the ratio of the expected variance to the expected mean. For a set number of independent variables, the smaller the dispersion multiplier, the better. For example, in a pure Poisson distribution, the dispersion should be 1.0. In practice, a ratio greater than 10 indicates that there is too much variation that is unaccounted for in the model. Either add more variables or change the functional form of the model.
28. Z-test for dispersion multiplier (Poisson models only). This is a test for whether the dispersion parameter is significantly greater than that assumed by the Poisson model. It is a test of over-dispersion.
29. P-value for Z-test of dispersion parameter (Poisson models only). This is the one-tail probability level associated with the Z-test.
30. Inverse dispersion multiplier. For a set number of independent variables, a larger inverse dispersion multiplier is better. A ratio close to 1.0 is considered good.

MLE Individual Coefficient Statistics

For the individual coefficients, the following are output:

31. The coefficient. This is the estimated value of the coefficient from the maximum likelihood estimate.
32. Standard Error. This is the estimated standard error from the maximum likelihood estimate.
33. Pseudo-tolerance. This is the tolerance value based on a normal prediction of the variable by the other independent variables.
34. Z-value. This is asymptotic Z-test that is defined based on the coefficient and standard error. It is defined as Coefficient/Standard Error.
35. p-value. This is the two-tail probability level associated with the Z-test.

Table 20.1 show the output for an MLE Poisson-Gamma model that relates the number of Houston 2007-09 burglaries to the number of 2008 households and the 2000 median household income of Traffic Analysis Zones.

Table 20.1:
Maximum Likelihood Output for Poisson-Gamma Model

```
 Model result:
 Data file:                          Burglaries_within_City_of_Houston.dbf
 DepVar:                             BURG2006
 N:                                  1179
 Df:                                 1175
 Type of regression model:           Poisson-Gamma-no spatial autocorrelation
 Method of estimation:               MLE

      Likelihood statistics
 Log-likelihood:                     -4430.800180
 AIC:                                8869.600361
 BIC/SC:                             8889.890048
 Deviance:                           1390.149554 P-value of Deviance:   0.0001
 Pearson Chi-Square:                 1112.717355 P-value of Chi-Square: 0.0001

      Model error estimates
 Mean absolute deviation:            39.580568
      1st (highest) quartile:        124.121350
      2nd quartile:                  19.377810
      3rd quartile:                  6.195620
      4th (lowest) quartile:         8.940150
 Mean squared predicted error:       62031.156586
      1st (highest) quartile:        242037.095867
      2nd quartile:                  6445.778853
      3rd quartile:                  118.261739
      4th (lowest) quartile:         154.880457

      Dispersion tests
 Adjusted deviance:                  1.183106  P-value of Deviance:   n.s.
 Adjusted Pearson Chi-Square:        0.946993  P-value of Chi-Square: n.s.
 Dispersion multiplier:              1.534057  Z= 910.799548 P-value: 0.0001
 Inverse dispersion multiplier:      0.651866

------------------------------------------------------------------------------
                                          Pseudo-
 Predictor   DF  Coefficient  Stand Error   Tolerance   z-value      p-value
 INTERCEPT   1     2.321019     0.083077           .    27.938042    0.001
    HH2006   1     0.001160     0.000066     0.993563   17.661356    0.001
MEDHHINC00   1    -0.000008     0.000002     0.993563   -5.129752    0.001
------------------------------------------------------------------------------
```

Markov Chain Monte Carlo (MCMC) Model Output

The MCMC routines (Normal-CAR/SAR, Poisson-Gamma, Poisson-Gamma-CAR/SAR, Poisson-Lognormal, Poisson-Lognormal-CAR/SAR, Binomial Logit, Binomial Logit-CAR/SAR) produce a standard output and an optional expanded output**.** The standard output includes summary statistics and estimates for the individual coefficients. Background information on these models is found in chapters 16, 17, 18, and 19.

MCMC Summary Statistics

The summary statistics include:

Information About the Model

1. The dependent variable
2. The number of records
   The sample number. This is only output when the block sampling method is used.
3. The number of cases for the sample. This is only output when the block sampling method is used.
4. Date and time for sample. This is only output when the block sampling method is used
5. The degrees of freedom (N – number of parameters estimated)
6. The type of regression model (Normal/OLS, Poisson, Poisson with linear correction, Poisson-Gamma, Poisson-Gamma-CAR/SAR, Poisson-Lognormal, Poisson-Lognormal-CAR/SAR, Binomial Logit, Binomial Logit-CAR/SAR)
7. The method of estimation
8. The number of iterations
9. The 'burn in' period
10. The block size is the expected number of records selected for each block sample. The actual number may vary.
11. The number of samples drawn. This is output when the block sampling method used.
12. The average block size. This is output when the block sampling method used.
13. The type of distance decay function used. This is output for models that use CAR or SAR spatial autocorrelation functions.
14. Condition number for the distance matrix. If the condition number is large, then the model may not have properly converged. This is output for the Poisson-Gamma-CAR model only.

15. Condition number for the inverse distance matrix. If the condition number is large, then the model may not have properly converged. This is output for the Poisson-Gamma-CAR/SAR, or Poisson-Lognormal-CAR/SAR models only.

Likelihood Statistics

16. Log-likelihood estimate, which is a negative number. For a set number of independent variables, the smaller the log-likelihood (i.e., the most negative) the better.
17. Log-likelihood per case. This divides the log-likelihood by the sample size (N). This indicates the average contribution to the log-likelihood of each observation. The more negative, the better.
18. Deviance Information Criterion (DIC) for models only. This adjusts the log-likelihood for the effective degrees of freedom. The smaller the DIC, the better.
19. Akaike Information Criterion (AIC) adjusts the log-likelihood for the degrees of freedom. The smaller the AIC, the better.
20. AIC per case. This divides the AIC statistic by the sample size (N). This indicates the average contribution to the AIC of each observation. The smaller, the better.
21. Bayesian Information Criterion (BIC), sometimes known as the Schwartz Criterion (SC), adjusts the log-likelihood for the degrees of freedom. The smaller the BIC, the better.
22. BIC per case. This divides the BIC/SC statistic by the sample size (N). This indicates the average contribution to the BIC/SC of each observation. The smaller, the better.
23. Deviance compares the log-likelihood of the model to the log-likelihood of a model that fits the data perfectly. A smaller deviance is better.
24. The probability value of the deviance based on a Chi-square test with *N-K-1* degrees of freedom where *K* is the number of independent variables.
25. Pearson Chi-square is a test of how closely the predicted model fits the data. A smaller Chi-square is better since it indicates the model fits the data well.
26. The probability value of the Pearson Chi-square based on a Chi-square test with *N-K-1* degrees of freedom where *K* is the number of independent variables.

Model Error Estimates

27. Mean Absolute Deviation (MAD). For a set number of independent variables, a smaller MAD is better.
28. Quartiles for the Mean Absolute Deviation. For any one quartile, smaller is better.

29. Mean Squared Predictive Error (MSPE). For a set number of independent variables, a smaller MSPE is better.
30. Quartiles for the Mean Squared Predictive Error. For any one quartile, smaller is better.

Dispersion Tests

31. Adjusted deviance. This is a measure of the difference between the observed and predicted values (the residual error) adjusted for degrees of freedom. The smaller the adjusted deviance, the better. A value greater than 1 indicates over-dispersion.
32. The probability value of the adjusted deviance based on a Chi-square test with 1 degree of freedom.
33. Adjusted Pearson Chi-square. This is the Pearson Chi-square adjusted for degrees of freedom. The smaller the Pearson Chi-square, the better. A value greater than 1 indicates over-dispersion.
34. The probability value of the adjusted Pearson Chi-square based on a Chi-square test with 1 degree of freedom.
35. Dispersion multiplier. This is the ratio of the expected variance to the expected mean. For a set number of independent variables, the smaller the dispersion multiplier, the better. In a pure Poisson distribution, the dispersion should be 1.0. In practice, a ratio greater than 10 indicates that there is too much variation that is unaccounted for in the model. Either add more variables or change the functional form of the model.
36. Inverse dispersion multiplier. For a set number of independent variables, a larger inverse dispersion multiplier is better. A ratio close to 1.0 is considered good.

MCMC Individual Coefficient Statistics

For the individual coefficients, the following are output:

37. The mean coefficient. This is the mean parameter value for the *N-K* iterations where *k* is the 'burn in' samples that are discarded. With the MCMC block sampling method, this is the mean of the mean coefficients for all block samples.
38. The standard deviation of the coefficient. This is an estimate of the standard error of the parameter for the *N-K* iterations where *k* is the 'burn in' samples that are discarded. With the MCMC block sampling method, this is the mean of the standard deviations for all block samples.
39. t-value. This is the t-value based on the mean coefficient and the standard deviation. It is defined by Mean/Std.

40. p-value. This is the two-tail probability level associated with the t-test.
41. Adjusted standard error (Adj. Std). The block sampling method will produce substantial variation in the mean standard deviation, which is used to estimate the standard error. Consequently, the standard error will be too large. An approximation is made by multiplying the estimated standard deviation by $\sqrt{\frac{\bar{n}}{N}}$ where $\bar{n}$ is the average sample size of the block samples and *N* is the number of records. If no block samples are taken, then this statistic is not calculated.
42. Adjusted t-value. This is the t-value based on the mean coefficient and the adjusted standard deviation. It is defined by Mean/Adj_Std. If no block samples are taken, then this statistic is not calculated.
43. Adjusted p-value. This is the two-tail probability level associated with the adjusted t-value. If no block samples are taken, then this statistic is not calculated.
44. MC error is a Monte Carlo simulation error. It is a comparison of the means of *m* individual chains relative to the mean of the entire chain. By itself, it has little meaning.
45. MC error/Std is the MC error divided by the standard deviation. If this ratio is less than .05, then it is a good indicator that the posterior distribution has converged.
46. G-R stat is the Gelman-Rubin statistic which compares the variance of *m* individual chains relative to the variance of the entire chain. If the G-R statistic is under 1.2, then the posterior distribution is commonly considered to have converged.
47. Spatial autocorrelation term (Phi, ϕ) for CAR/SAR models only. This is the estimate of the fixed effect spatial autocorrelation effect. It is made up of three components: a global component (Rho, ρ); a local component (Tauphi, τ_φ); and a local neighborhood component (Alpha, α, which is defined by the user).
48. The log of the error in the model (Taupsi). This is an estimate of the unexplained variance remaining. Taupsi is the exponent of the dispersion multiplier, $e^{\tau\psi}$. For any fixed number of independent variables, the smaller the Taupsi, the better.

Expanded Output (MCMC Only)

If the expanded output box is checked, additional information on the percentiles from the MCMC sample are displayed. If the block sampling method is used, the percentiles are the means of all block samples. The percentiles are:

49. 2.5^{th} percentile
50. 5^{th} percentile
51. 10^{th} percentile

52. 25th percentile
53. 50th percentile (median)
54. 75th percentile
55. 90th percentile
56. 95th percentile
57. 97.5th percentile

The percentiles can be used to construct confidence intervals around the mean estimates or to provide a non-parametric estimate of significance as an alternative to the estimated t-value in the standard output. For example, the 2.5th and 97.5th percentiles provide approximate 95 percent confidence intervals around the mean coefficient while the 0.5th and 99.5th percentiles provide approximate 99 percent confidence intervals.

The percentiles will be output for all estimated parameters including the intercept, each individual predictor variable, the spatial effects variable (Phi), the estimated components of the spatial effects (Rho and Tauphi), and the overall error term (Taupsi).

Table 20.2 show selective output from an MCMC Poisson-Lognormal-CAR spatial model that relates the number of Houston 2007-09 burglaries to the number of 2008 households and the 2000 median household income of Traffic Analysis Zones. The percentiles have been reduced to 0.5th, 2.5th, 97.5th, and 99.5th to fit the table.

Output Phi Values (CAR/SAR Models Only)

For the CAR and SAR models only, the individual Phi values can be output. This will occur if the sample size is smaller than the block sampling threshold. Check the 'Output Phi value if sample size smaller than block sampling threshold' box. An ID variable must be identified and a DBF output file defined.

Save Output

The predicted values and the residual errors can be output to a 'dbf' file with a REGOUT<*root name*> file name where *rootname* is the name specified by the user. The output is saved under a different file name. The output includes all the variables in the input data set plus two new ones: 1) the predicted values of the dependent variable for each observation (with the field name PREDICTED); and 2) the residual error values, representing the difference between the actual /observed values for each observation and the predicted values (with the field name RESIDUAL). The file can be imported into a spreadsheet or graphics program and the errors plotted against the predicted dependent variable (similar to Figure 15.3 in chapter 15).

Table 20.2:
MCMC Output for Poisson-Lognormal-CAR Model

```
DepVar:                          BURG2006
 N:                               1179
 Df:                              1174
 Number of iterations:            25000
 Type of regression model:        Poisson-Lognormal-CAR
 Method of estimation:            MCMC
 Distance decay function:         Negative exponential

      Likelihood statistics
 Log-likelihood:                  -6087.822981
      Per case:                   -5.163548
 DIC:                             30510.458212
 AIC:                             12185.645963
      Per case:                   6.246823
 BIC/SC:                          7390.366951
      Per case:                   6.268335
 Deviance:                        414.787381     P-value of Deviance:   0.0001
 Pearson Chi-Square:              422.236291     P-value of Chi-Square: 0.0001

      Model error estimates
 Mean absolute deviation:         5.387914
      1st (highest) quartile:     14.262519
      2nd quartile:               5.504652
      3rd quartile:               1.340483
      4th (lowest) quartile:      0.493941
 Mean squared predicted error:    149.000118
      1st (highest) quartile:     542.211088
      2nd quartile:               51.172821
      3rd quartile:               3.835512
      4th (lowest) quartile:      0.298416

      Dispersion tests
 Adjusted deviance:               4.456926     P-value of Deviance:      0.0001
 Adjusted Pearson Chi-Square:     20.611149    P-value of Chi-Square:    0.0001
 Dispersion multiplier:           0.904852     Z = 133.050700            P-value of Z: 0.0001
 Inverse dispersion multiplier:   1.105154
```

	Mean	Std	t-value	p-value	MC error	MC error/ std	G-R stat
Intercept:	0.057768	0.086334	0.669124	n.s.	0.001960	0.022698	1.004705
HH2006:	0.000156	0.000064	2.448304	0.02	2.7333e-006	0.042825	1.018906
MEDHHINC00:	-5.7411e-008	1.5194e-006	-0.037785	n.s.	2.7607e-008	0.018169	1.001817
Spatial autocorrelation (Phi):	1.660699	0.063369	26.206660	0.001	0.003377	0.053283	1.026494
Global component (Rho)	0.178264	0.142500	1.250969	n.s.	0.001356	0.009515	1.000174
Local component (Tauphi):	0.003762	0.000404	9.317862	0.001	0.000018	0.043832	1.019646
Neighborhood component (Alpha: defined)	-0.636652 Miles						

Table 20.2 (continued)

Percentiles	0.5th	2.5th	97.5th	99.5th
Intercept:	-0.153012	-0.103269	0.236579	0.300385
HH2006:	0.000008	0.000041	0.000289	0.000339
MEDHHINC00:	-0.000004	-0.000003	0.000003	0.000004
Spatial component				
(Phi):	1.486406	1.530011	1.776679	1.804173
Global component				
(Rho):	0.001125	0.005925	0.525452	0.657165
Local component				
(Tauphi):	0.002892	0.003060	0.004649	0.005008

Save Estimated Coefficients

The individual coefficients can be output to a DBF file with a REGCOEFF<*root name*> file name where *rootname* is the name specified by the user. This file can be used in the 'Make Prediction' routine under Regression II.

Diagnostics Relevant for Spatial Regression

In chapter 15, the diagnostic tests for the regression module were described. Among the statistics produced by the routine are two relevant for spatial regression.

Testing for Spatial Autocorrelation in the Dependent Variable

First, there is the Moran's "I" test for spatial autocorrelation. The statistic was discussed extensively in Chapter 5. If the "I" is significant, *CrimeStat* outputs a message indicating that there is definite spatial autocorrelation in the dependent variable and that it needs to be accounted for, either by a proxy variable or by estimating a CAR or SAR model.

A *proxy* variable would be one that can capture a substantial amount of the primary reason for the spatial autocorrelation. One such variable that we have found to be very useful is the distance of the location from the metropolitan center (e.g., downtown). Almost always, population densities are much higher in the central city than in the suburbs, and this differential in density applies to most phenomena including crime (e.g., population density, employment density, traffic density, events of all types). It represents a *first-order* spatial effect, which was discussed in Chapters 4 and 5, and is the result of other processes. Another proxy variable that can be used is income (e.g., median household income, median individual income) which tends to account for much clustering in an urban area. The problem with income as a proxy variable is that it is both causative (income determines spatial location) as well as a by-product of

population densities. The combination of both income and distance from the metropolitan center can capture most of the effect of spatial autocorrelation.

An alternative is to use the Poisson-Gamma-CAR model to filter out some of the spatial autocorrelation. As we discussed above, this is useful only when all obvious spatial effects have already been incorporated into the model. A significant spatial effect only means that the model cannot explain the additional clustering of the dependent variable.

Estimating the Value of Alpha (α) for CAR/SAR Models

Second, there is an estimate of a plausible value for the distance decay function alpha, α, in the CAR or SAR models. The way the estimate is produced was discussed above and is based on assigning a proportional weight for the distance associated with the nearest neighbor distance, the average distance from each observation to its nearest 'neighbor' (see chapter 6).

Three values of α are given in different distance units, one associated with a weight of 0.9 (a very steep distance decay, one associated with a weight of 0.75 (a moderate distance decay), and one associated with a weight of 0.5 (a shallow distance decay). Users should run the Moran Correlogram and examine the graph of the drop off in spatial autocorrelation to assess what type of decay function most likely exists. The user should choose an α value that best represents the distance decay and should define the distance units for it.

Regression II Module

The Regression II module allows the user to apply a model to another dataset and make a prediction. Figure 20.3 show the Regression II setup page. The 'Make prediction' routine allows the application of coefficients to a dataset.

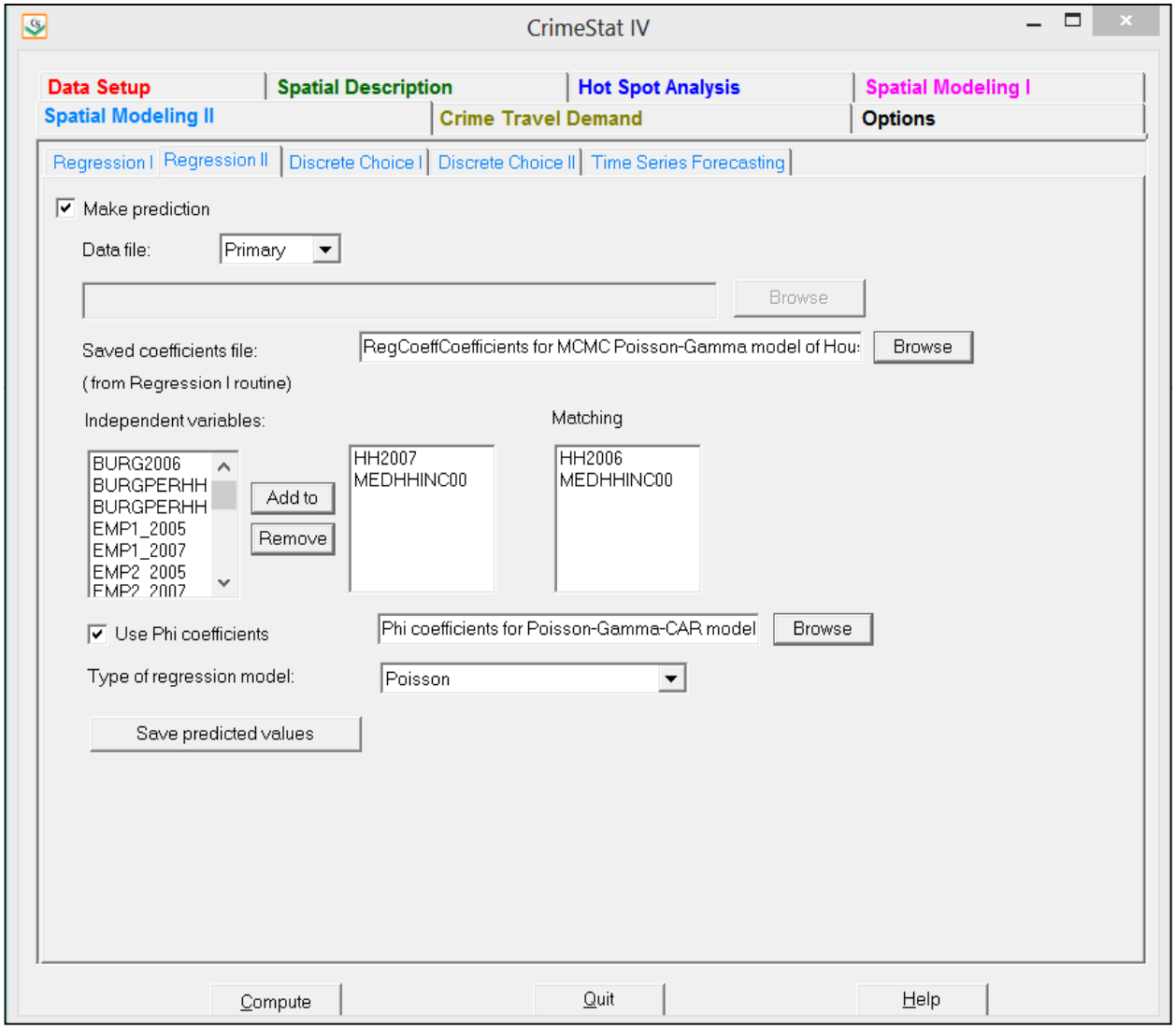
Note that, in this case, the coefficients are being applied to a different Primary File than that from which they were calculated. For example, a model might be calculated that predicts robberies for 2006. The saved coefficient file then is applied to another dataset, for example robberies for 2007.

There are four types of models that are fitted – normal, Poisson, binomial logit, and binomial probit. For the normal model, the routine fits the equation:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} \qquad (20.9)$$

Figure 20.3:

Regression Modeling II Setup Screen



For the Poisson model, the routine fits the equation:

$$Y_i = e^{\beta_0+\beta_1 X_{1i}+\cdots+\beta_K X_{ki}+[\Phi_i]} \tag{20.10}$$

with β_0 being the intercept (if calculated), $\beta_1 \ldots \beta_k$ being the saved coefficients and Φ_i is the saved Phi values (if a CAR or SAR model was estimated). Notice that there is no error in each equation. Error was part of the estimation model. What were saved were only the coefficients.

For the binomial logit model, the routine fits the equation:

$$P(Y=1) = \frac{e^{\beta_0+\sum_1^K \beta_K X_K+[\Phi_i]}}{1+e^{\beta_0+\sum_1^K \beta_K X_K+[\Phi_i]}} = \frac{1}{1+e^{-(\beta_0+\sum_1^K \beta_I X_K+[\Phi_i])}} \tag{20.11}$$

with β_0 being the intercept (if calculated), $\beta_1 \ldots \beta_k$ being the saved coefficients and Φ_i is the saved Phi values (if a CAR or SAR model was estimated).

For the binomial probit model, the routine fits the equation:

$$p(Y=1) = \Phi^{-1}(p_i) = \beta_0 + \sum_1^K \beta_K X_K \tag{20.12}$$

with β_0 being the intercept (if calculated), $\beta_1 \ldots \beta_k$ being the saved coefficients and Φ_i is the saved Phi values (if a CAR or SAR model was estimated), and Φ is the cumulative standard normal distribution,

For all four types of model, the coefficients file must include information on the intercept and each of the coefficients. The user reads in the saved coefficient file and matches the variables to those in the new dataset based on the order of the coefficients file.

If the model had estimated a general spatial effect from a CAR or SAR model, then the general Φ_i will have been saved with the coefficient files. If the model had estimated specific spatial effects from a CAR or SAR model, then the specific Φ_i values will have been saved in a separate Phi coefficients file. In the latter case, the user must read in the Phi (Φ_i) coefficients file along with the general coefficient file.

Table 20.3 shows the output for the first 20 cases from a prediction of the number of burglaries per zone based on the estimation model shown in Table 20.2 (Poisson-Lognormal-CAR). The output will include all variables in the input data set plus the Phi coefficient and the predicted values. The user can then calculate residuals by subtracting the predicted from the actual (observed) values of the dependent variable.

Table 20.3:
File Output from Poisson-Lognormal-CAR Prediction of Houston Burglaries

TAZ03	BURG2006	PHI	PREDICTED
532	19	0.633593	7.922792
534	2	-0.163279	7.030844
536	2	-0.223977	11.555803
530	107	1.323602	21.462356
537	19	0.259453	15.658255
522	55	1.537228	5.987060
538	11	0.335330	7.432503
516	10	0.364732	9.598958
481	0	-0.350902	8.693915
474	1	-0.161788	8.348133
482	7	0.009940	12.178501
496	2	-0.245535	17.342402
548	0	-1.199179	13.717904
475	4	-0.037407	8.166218
435	3	-0.425498	8.307698
476	1	-0.056756	8.897897
484	8	0.014615	16.133065
483	2	-0.066611	7.888521
477	1	-0.076599	8.166218
478	0	-0.050627	9.293352

Conclusion

This chapter has summarized the structure of the Regression I and Regression II modules and most of the options that are available. The help menu on the program will provide context-specific help on individual items. Note that if you are using Windows Vista, Windows 7 or Windows 8, you must download a utility from *Microsoft* that allows the help menu to be viewed from the program. See Chapter 1 (p. 1.17) for details.